7.5.67
#4

Title Page

THE DIGITAL COMPUTER
AS A TOOL
OF NEW TESTAMENT STUDY

BY

CHARLES EDWARD CURTIS

What do they use at the Holy Ghost dentist's office? cf p 23

Submitted in partial fulfillment of the requirements
for the Bachelor of Divinity degree
in the Episcopal Theological School
June, 1967

TABLE OF CONTENTS

EXPLANATORY NOTE

The location of reference to some of the particular statistical studies and their applications is as follows:

> Yule's use of 'K' characteristic (pp. 36, 39-40 of the text) can be located according to p. 36, n. 41; Somers' use of the Discriminant of Fisher, the T^2 test of Hotelling, Type-Token Ratio, his own Θ measurement, and Factor Analysis (pp. 70-72 of the text) can be located according to p. 70, n. 66.

The pertinent notes on the material near the mention of these studies in the text will help in more precise location.

CHAPTER I

INTRODUCTION

The digital computer as a tool of New Testament study is not easily defined. It involves several fields of work which are not always related. Computer engineering statistics, literary stylistic analysis, word cataloging, and concordance making, New Testament criticism, semantic analysis, and literary criticism are all involved in one way or another. This is not a simple, singular involvement, for some are bound up with two or three of the others. The use of the computer in lexical and linguistic analysis and in other work with the New Testament can not be seen apart from the whole growth of computer-oriented analyses and the growth in the application of statistics to literary considerations in general. In a sense, the separation of Chapters 2 and 3 represents a false dichotomy. They are both part of one major movement. The break is demanded by the particular special interest which prompted this writing. Chapter 4 represents the logical outgrowth of the work that has been done so dar. It unites the work in Chapters 2 and 3 in pursuing the further usefulness of the computer in New Testament study.

3
2
1
0

The Nature of the Beast

To help in the understanding of the utility of the digital computer in work with the New Testament some discussion of the nature and function of digital computers is necessary. The digital computer is an electronic machine which has been conceived and developed within the last twenty-five years. Its principle of operation is based on the use of two discrete stable states - "on" and "off." From this basis, an internal system of binary numerical functions are constructed in such a way that addition can be easily accomplished. Binary mathematics (with "2" as a base instead of the more usual decimal system's "10") is the most useful expression of the on-off function ("0" and "1" are the two numbers) and also the simplest to construct using the on-off basis of operations. With the speed of electronic equipment the digital computer can perform several other functions using addition as the basis of operation. Addition with negative numbers constitutes subtraction. Multiplication is accomplished by addition of one number (the "multiplicand") to itself a certain number of times (the "multiplier"). Similarly, division is performed through addition of the divisor in a negative state to the dividend with the machine adding the number of times the negative addition (subtraction) is made, the answer being the quotient.

The digital computer differs from its associate, the analog computer, in that the latter is based on the relationship of physical quality to numerical value, but the former is based on discrete numerical functions. Common examples of the principle of analog computers are graphs and sliderules. The abacus, on the other hand, uses the principle applied by the digital computer. The digital computer is thus accurate in a way that analog computers cannot be. By using on-off it can calculate to any number of figures limited only by the size of the machines ability to remember the numbers. This is in marked contrast to the necessity, with an analog computer, of "rounding off" the answer. After a time it becomes mere guesswork as to exactly what the quantity is. The analog computer is capable of giving dependable answers only up to e.g. five digits, irrespective of decimal point location (12579.0 or .0012579). After that the next digit must be guessed at based on experience and a sharp eye. The data to be used in the functioning of the digital computer are called "input." The input must be in a form which the machine can use and work with. The most commonly seen form of input is the punch card. The letters and numbers are punched in different arrangements of twelve possible punch places in each of eighty columns on the card. In a card reader, the machine translates the holes into binary numbers with which to work. Paper punched tapes are another form of

input. It uses a paper tape on which the possible combinations of a five-hole column are punched to enable the machine to read the punches and translate them into internal computer binary mathematics. The use of magnetic tape and magnetic disks are high speed mechanisms which utilize the same type of format as the punch cards. However, due to the magnetic rather than mechanical nature of their coding, they are able to be processed and scanned at a much faster rate than either punch cards or paper tape.

Once in the computer, the input is processed according to a "program." A program is a set of instructions which can be written in any one of a number of statistical languages designed to be read and understood by the computer. The part of the computer called a "complier" compiles the machine's instructions by translating the program language into machine language - binary mathematics. The information in the machine is stored in "storage" or "memory." It is stored as electronic impulses which are held by magnetized pieces of metal, one binary digit to a piece. They can be recalled from storage for calculation and returned.

In giving an answer in a comprehensible fashion, the computer is limited only by the limits of the requirements for different forms of the "output." The output can be in the form of punch cards, paper tape, magnetized tape or disks, ready to be used again in the computer or processed into another form. High speed printers can print the results,

as can specially prepared typewriters. Graphs and visual displays by cathode-ray tube are also available. Any and all of these output forms may be utilized in producing the results of the computer's action in a comprehensible form.

The Powers of the Beast

The utility of the computer can be sharply defined. The computer can perform any number of mechanical functions on any data that it can convert to binary numbers. It can do nothing more. It can perform mathematical and analytical functions on numbers, letters, or symbols which are put in. This is a definite advantage, but it is also a disadvantage. Since the computer functions logically and mechanically, it will not do anything that it is not told to do and will do everything that it is told to do without variation. Since Howard Aiken of Harvard opened this field in 1944 with Mark I[1] the limits within which the computer can work have been greatly expanded. However, the framework of logical, mathematical actions remains. Le Corbeiller pointed out that the computer solves problems when the solutions are "uniquely determined by the data."[2] This left two kinds of problems which the machine could not solve at all. The

---

[1] Le Corbeiller, Philippe, "What We Should Learn from Computers," in Proceedings of a Harvard Symposium on Digital Computers and Their Applications, p. 1.

[2] Ibid., p. 5.

computer is helpless in the face of a lack of data or of contradictory data. To complete the logical solution of a problem the computer must have all of the data necessary to the solution and none must be contradictory. Neither can the computer handle problems that call for value judgments. This is not a logical sequence within the framework of the computer - it is not made wholly on the basis of the data - and the computer thus cannot undertake to perform it. (The machine may choose X over Y and will do so consistently, if it is told under what precise circumstances to do so.) In such a case where the value judgment is not programmed into the computer as a logical operation (if X, then Y), the machine returns the problem with no solution or with infinite solutions.[3]

Paul Tasman's article on the processing of literary data noted the advantages which the computer offers to literary data processing.[4] Acceleration of study, the ability to explore the text more rapidly, and the speed of alphabetizing both from left to right and from right to left (especially useful for inflections and Hebrew) were discussed. An interesting comparision in this regard is the estimation of work needed to index and concordize 2000 pages from the Summa Theologica of St. Thomas, amounting to nearly

---

[3] Ibid., pp. 5-6.

[4] Tasman, Paul, "Literary Data Processing," in IBM Journal of Research and Development 1:254, July, 1957.

1.6 million words. Using the old manual methods it would take 3 persons 20,000 hours. Using conventional punch cards and sorters it would take 3 persons 1,000 hours. Using large scale computers it would take one person 60 hours, exclusive of preparation and programming which could be used for all data to be processed in this manner.[5]

On the whole, the lexical work, using the computer for the study and categorizing of words as entities, has tended to follow the pattern set by the old "3 x 5 card handling techniques."[6] Being thorough, logical, and generally accurate[7] the computer can do these functions better than men. Computers do not take into account the significance of a variation from normal or the value of including it in a set of figures. They only record it and evaluate it equally with similar occurrences. This is especially valuable on longer projects or projects involving larger counts. In Mosteller and Wallace's work[8] the words were counted both by hand and by machine. The benefits of computers can be seen from this recounting of the manual method:

---

[5] Ibid., p. 256.

[6] Heller, Jack, "A Proposed System for the Collection, Correction and Rearrangement of Large Masses of Data," in Proceedings, Literary Data Processing Conference, Yorktown Heights, N.Y., 1964, p. 98.

[7] Errors made by the machine independently are very infrequent.

[8] Mosteller, F., and Wallace, D.L., Inference and Disputed Authorship: The Federalist.

> "Certain Federalist papers were typed on roll paper (adding machine paper, one word to a line) and proofed. At the same time, the page number and line number were written opposite each word (for some, but not all papers). The words on the roll paper were then cut and sorted into alphabetical order. (During this operation a deep breath created a storm of confetti and a permanent enemy)."[9]

The advantage of using a computer to perform searching, clerical, and statistical operations on literary data is especially impressed upon those who have done so. However, the limitations should not be overlooked in this application of mechanical analysis to literary data. As mentioned above, the computer performs only those functions for which it is equipped with a full set of non-contradictory instructions and data. The data which the computer is given, as well as the program which instructs it in what to do with the data, are both humanly conceived and humanly executed. The computer is the tool of men and at their command as to how to operate and with what to operate. The deleterious effect which this has on speed and accuracy should be noted here, and it will be dealt with later in Chapter 2. Not only are the input and program humanly controlled, but the output is interpreted by humans. The computer cannot "prove" anything; people can try to "prove" things using the computer. Properly controlled, the computer can prove that "2 + 2 = 5". This does not prove that "2 + 2 = 5," but that someone was able to program a computer properly to get

---

[9] Ibid., p. 44.

that result. This will be further discussed as this work moves on.

Literary Statistics

In discussing the application of computer techniques to the study of the New Testament, a treatment of the relationship of statistics to literary research is also necessary. As the computer has been successfully applied to the lexical problems of the New Testament, and of literature in general (e.g. indices, concordances, and textual comparisions), so it has also been applied to the analysis of literary data for such things as style and structure. This latter has been linked with the use of statistics for assistence both in the determination of the objective criteria of style and structure as well as in the interpretation of results coming from such studies.

The theoretical integration of the statistics of literary data and the field of linguistics was undertaken by Gustav Herdan.[10] His work serves not only as an introduction to this area, but is a rather extensive application of statistics to the whole field of linguistics.

In the linguistic analysis of the New Testament (which will be discussed at greater length in Chapter 3) the critical use of statistics is found ~~not~~ most in the analysis of

---

[10] Language as Choice and Chance, cf. p. 2.

style for use in the determination of authorship. The particular area of statistics used here is general discrimination or classification. The problem is to establish some criteria by which writings of different authors may be separated and authorship of doubtful works settled.

Among the problems faced is the generalized nature of statistics. Statistics provide greater detail and greater validity in dealing with groups and generalization. Mosteller and Wallace noted this in dealing with the Federalist papers. The results of the initial surveys indicated a general trend toward Madison in the disputed papers. However, this could not be used to settle individual essays satisfactorily.[11] The problem of the reliability of the statistical results must be faced especially when dealing with the New Testament. A set of samples, adequate in size, of two different authors between whom the disputed work is to be decided is the statistical requirement for greatest reliability in the use of statistical criteria. In the New Testament studies the size of the samples is limited by the size of the writings, and the writings which are disputed can often be compared only among themselves. This tends to lower the reliability of the statistical analysis in giving definite conclusive results.

---

[11] Mosteller, F., and Wallace, D.L., "Notes on an Authorship Problem," in Proceedings of a Harvard Symposium on Digital Computers and Their Application, p. 167.

Computers and Literary Study

The attitude of scholars in literary studies toward the use of the computer has been categorized into three groups.[12] The first group is hostile, basing its attitude on the supposed incompatibility of humanities and automation. This group is "threatened" by the use of computers and refuses to conceive of its valid application in the research of literary studies.[13] The second group, believing in the omnipotence of the machine, are more dangerous because their attitude "leads to oversimplification of the problems involved, forgetting that a machine permits neither sloppy thinking nor mistakes."[14] The third group is critical in its use of the machine, not fearing it, but not blind to its limitations.

It is this last group which can most freely and most effectively use the computer to further research. The computer does perform certain tasks and perform them well, as has been noted above. However, it also demands strict attention to its abilities and limitations for the most valid use of what it does. The first group cited above has its counterpart in the emotions of the people using the computers themselves. J.B. Bessinger noted the psychological problems

---

[12] De Tollenaere, F., Nieuwe Wegen in de Lexicologie, pp. 139-140.

[13] Ibid., p. 139.

[14] Ibid., p. 140.

encountered when preparing a computer concordance.[15] He recounts an emotional reaction against "Instant Concordances" in the light of the time and effort that went into their making before the use of the computer. It did not seem honest to get similar results in such short time.[16] Stephen Parrish also echoes this as he notes some ambivalence in every good humanist about technological intrusion into his domain.[17] This is compounded by rumors of excessive pronouncements and wild claims for the abilities of the computer which humanists ignorantly accept, and then they raise their defenses against the "thinking machine" even higher.[18] This is part of the problem of a lack of communication between the humanists and computer engineers. This problem must be met before fuller exploration of machines for research in the humanities can be effected.

Part of that problem is found in uncertainty of the qualitative-quantitative barrier which exists in humanistic

---

[15] Bessinger, J.B., "Computer Techniques for an Old English Concordance," in American Documentation 12:227, July, 1961.

[16] Ibid., p. 227.

[17] Parrish, S.M., "Problems in the Making of Computer Concordances," in Studies in Bibliography 15:1, 1962.

[18] Ibid.,p. 2.

studies.[19] The relationship of judgment and measurement must be in some sense decided before the place of the computer in literary studies can be established and settled. Are the two - measurement and criticism - set in the sharp opposition of quantitative and qualitative as Louis Milic suggests?[20] Or does one lead into the other, measurement into criticism, as Alan Markham suggests?[21] This question is still not settled. However, in learning to use the computer and seeking its valid application to the problem of humanistic research, the exceptional ultimacy with which the question is vested in regards to the computer can be minimized. The computer can be valuably and successfully applied to many of the current problems in literary studies without a final answer to the quantitative-qualitative question. Perhaps the further pursuit of computer utilization will be itself more informative to that question.

The functioning of a computer in regards to human involvement in its operations is put quite well by Robert

---

[19] Parrish, S.M., "Summary," in Proceedings, Literary Data Processing Conference, Yorktown Heights, N.Y., 1964, p. 6.

[20] Milic, Louis, "Some Risks of Technological Overindulgence for the Humanities," in Proceedings, Literary Data Processing Conference, Yorktown Heights, N.Y., 1964, pp. 55-63.

[21] Markham, Alan, "Litterae ex Machina," in Proceedings, Literary Data Processing Conference, Yorktown Heights, N.Y., 1964, pp. 37-54.

Wachel. When the machine seems to think, it is because "some human agent with insight, imagination, ingenuity, and a great amount of time has first determined a completely specified procedure...for doing part of a complex job."[22] The machine "communicates these results with insight, imagination, ingenuity, and if he has done his original job well, in a small amount of time."[23] The machine works to do the drudgery, the common-place checking, filing, figuring, and calculating which the human could do with less accuracy and much more time. The scholar must still know what he is looking for, how to get it, and what to do with it when he gets it. The scholar is responsible for what he puts into the machine, what he gets out (although he may not have expected exactly what he received), and what he does with what he got out. If any of these are lacking in accuracy or integrity, the whole process will be of low quality. The computer is a constant, the scholars and researchers who use it determine the reliability which can be placed in the results of its labor.

---

[22] Wachel, Robert, "On Using a Computer," in The Computer and Literary Style, edited by Jacob Leed, p. 14.

[23] Ibid., p. 14.

CHAPTER II

THE COMPUTER IN LITERARY STUDIES

This chapter provides the background against which is set the use of the computer in New Testament studies. The relevant fields of literary data processing are reviewed and discussed.

Literary data processing can be broken down into the two general headings of lexical and supralexical processes. By lexical is meant those processes which deal with the words in seemingly clerical fashion. Concordance building and textual collation are the most obvious examples of this. Mechanical translation also falls in this category because the machine is essentially involved in an exchange process for this type of program. A word is recognized and changed to another expression and printed. (Linguistic structural study is necessary in assisting this, but it is not a part of the translation itself.)

Under supralexical processes fall linguistic and semantic literary analysis. Linguistic processes abstract the language patterns from the text through examining the text for those patterns. While still essentially a clerical function the object of the study is not word manipulation but the seeking of structure. Semantic literary analysis, likewise, seeks to detect and abstract the semantic patterns from the text. Presently this is most easily done in a coded form

(as are several linguistic studies such as sentence structure analysis). Whether it works from the text or from codes derived from the text the computer is an integral part of the processes of linguistic and semantic literary analysis.

The place of computers in mechanical translation and concordance and textual study will be quite apparent in the discussion that follows. In dealing with linguistic literary analysis the computer has seemed to become a bit sidetracked. The introduction of statistical considerations, especially sophisticated statistics, to literary analysis is quite recent. To understand the place of the computer in literary analysis it is necessary to see it through the parallel development of statistical application to literary studies. The computer has not been so integrally connected to linguistic literary analysis as it has been to mechanical translation or concordance compilation. Rather, it has been used as the servant of statistics, collecting the data, compiling the figures, and printing out the results for the statistical demands of linguistic analysis. If this section is viewed with this in mind the connection between the computer and linguistic literary analysis will be more apparent.

The use of the computer in the semantic literary analysis is of a similar nature. Attached to the method after its conception, the computer serves it by doing the counting and computing to make it more accurate and speedy in execution.

The limitations of input form, common to all of the literary data processing procedures by computer, is discussed in a separate section. It affects seriously the work which is being done by increasing the possibility of error, by slowing down the initiation of computer solution for the various problems and procedures, and by increasing the cost through added human steps in the processing of literary data.

A survey of the non-Biblical work in Greek text is added at the end. This is to provide a fuller background in the area with which the next chapter will deal -- the computer in the study of the New Testament. This work is shown to complete the picture of Greek literary studies as seen in the next chapter.

Mechanical Translation

One of the more glamorous applications of the computer to literary matters has been its adoption as a means (or a possible means) of producing adequate translations of written material from one language to another. From a start in 1946[1] this use of computers to translate has borne fruit after much hard work, although even more is left to do. In matter of process the computers translate following the same course as did (and do) their human counterparts. However, the

---

[1]Booth, A. D.; Brandwood, L.; and Cleave, J. P., Mechanical Resolution of Linguistic Problems, p. 1.

computer, due to its completely logical and mathematical nature, must have its process spelled out exactly for it. It is also able to deal only with the words which appear in its input, and not the ideas which the words represent, the translation of which is the important reason for attempting the translation at all. It is the successful solution of the problems raised by these limitations which makes the computer able to render an understandable and worthwhile translation.

The computer translates on a three-fold model.[2] The process in its simplest form is: Analysis, the coding of the input information; Conversion, the substitution of one code for another; and Synthesis, the changing of the new code to text in the output language. On the lowest level of this process translation occurs on a word for word basis, a direct, literal translation approximating the style of transliteration. This, however, produces an unintelligible and often humorous string of words which are not of any consistent value as a workable translation. The problem of differing frameworks is at the root of this particular liability. The language of input and the language of output are set in two different linguistic frameworks. Any attempt to change one to the other without adjusting the framework is doomed to failure.[3] The

---

[2]Delavenay, Emile, An Introduction to Machine Translation, p. 52.

[3]Ibid., p. 8.

attempts to solve this problem will be dealt with further below in this section.

There are also several blocks to an acceptable translation which are representative of problems met in other areas of literary computer work, and which result from the computer's singularly literal capabilities. These problems are the elementary incapability of the computer to distinguish homographs of varied meaning, to separate idioms from the same group of words with their basic, literal meaning, and to associate different members of an inflection with their root word. These were brought out in the construction of a bilingual dictionary within access of the computer.[4] The dictionary provides the direction for the substitution of input code for output code. However exact and precise the dictionary might be, it is in and of itself unable to provide distinction as to the various possible meanings, and thus translations, of a textual homograph of several different meanings. It is also unable in and of itself to distinguish the idiomatic meaning or the literal meaning of an idiom. (This is especially important when the idiom must be changed in the new language to remain constant in meaning.) The computer is not able to associate various inflections of a root when run on a simple dictionary language

[4]For a much more detailed treatment and explanation of the dictionary compilation see Oettinger, A. G., Automatic Language Translation, pp. 216-351.

converter program. While this drawback is more problematic in areas to be discussed below, it would still be helpful if several forms could be recognized as one in order to provide greater flexibility, especially with verbs.

At this point with these conditions the machine is of little practical use in aiding translation. The human effort is but slightly reduced in providing an adequate translation for desired texts. It then becomes necessary to provide adequate mechanical resolution for these problems to make computer translating of substantial quality and independence of constant human attention.

These problems have in large part actually been solved. In the problem of paradigmatic association a simple instruction to the computer to class the different inflections of one root together was sufficient. (However, a memory size limitation within the computer was, for a time, a hinderance to the successful pursuit of this.[5] This has been relieved by the larger storages of more modern computers.)

The problems involved with idioms were removed in much the same manner.[6] In the main stem dictionary an instruction for a word in a possible idiom sends the search to the idiom dictionary instead of simply translating it. The machine searches for the full idiom in the text which, if found, is

---

[5] Oettinger, op. cit., p. 340.

[6] Delavenay, op. cit., p. 89.

translated as the idiom, otherwise the machine returns to the main dictionary and translates the word ordinarily.

The problem of multiple meaning words, homographs, was less amenable to solution because the nature of the problem was entirely outside the analysis of words. (The different words could be grouped, and combinations of words given particular translations with recourse only to adjusting the machine's way of handling them. Since homographs are literally indistinguishable, they cannot be handled in this manner.) The homographic problem and the problem of disparate frameworks (above) are related because they both involve recourse to the syntactical and grammatical constructions within which the words are found.[7] This led researchers into new areas of literary analysis which resulted in the completion of subroutines which enabled the computer to use the context of the homographs to determine what meaning would be apropos, and thence to apply it. Delavenay points out that although this solves the problem in large part, there are still some homographs which defy this analysis. However, he goes on to say that this would be true even of human translators, and instead of an educated guess, the machine would simply print out the possibilities.[8]

---

[7] Ibid., pp. 54ff., 67ff.; c f. Yngve, Victor H., "Syntax and the Problem of Multiple Meaning", in Machine Translation of Language, edited by W. N. Locke and A. D. Booth, pp. 208-226.

[8] Delavenay, op. cit., pp. 90-91.

The linguist enters the process of machine translation through the setting up of frameworks for the translation. It is not within the machine's capacities, but within the linguist's, to set up within the machine the relationships of the systems of expression of input with those of output.[9] Through the human provision of inventories of expression frameworks for different languages and their means of conversion, the machine is then able to produce a more natural translation in the output language. (This specifically applies to such things as word order.)

Delavenay sees the way ahead for mechanical translation as optimistic. There does not seem to be any major stumbling block to progress in more detail.[10] The area which at present needs greatest development is the production of dictionaries. They are usually specialized by subject area , and the more precisely defined scientific subjects are receiving the greatest amount of attention presently. In the future, however, the dictionary compilation will move into the more 'general definition' areas of humanistic research and literary endeavors. With a greater portion of extralexical meanings associated with these area, e.g. metaphor, cliché, and dramatic situation, the programming becomes more complex and sophisticated. This will be where mechanical translation

---

[9] Ibid., pp. 45-46.

[10] Ibid., p. 123.

will be least effective. However, its usefulness in scientific text will be of increasing value for the speed with which it works and for its growing reliability in this more 'precise definitions' area of writing.

Concordances and Textual Study

Performing a function similar to that performed in automatic machine translation the computer has also been used for the production of concordances and in research into textual problems. (The contributions of John W. Ellison to both of these fields will be dealt with below in the next chapter as they are specifically related to the New Testament.) Among the earliest products of literary data processing,[11] concordances are also the product of non-linguistic functioning of the computer. The machine takes over the clerical task of arranging and rearranging text to form the alphabetic concordance.[12]

---

[11] Fogel, E. G., "The Humanist and the Computer: Vision and Actuality", in Proceedings, Literary Data Processing Conference, Yorktown Heights, N.Y., 1964, p. 16.

[12] A concordance may be described as in the title of John Marbecke's work published in 1550: A Concordance, that is to saie, a Worke wherein by the Ordre of the letters of the A. B. C. ye maie redely finde any Worde conteigned in the whole Bible so often as it is there expressed or mencioned. from Bessinger, J. B., "Computer Techniques for an Old English Concordance", in American Documentation 12:229, July, 1961.

A more 'relevant' explanation would be from Tasman, Paul, "Literary Data Processing", in IBM Journal of Research and Development 1:251, July, 1957: "A concordance is an alphabetical collection of the individual words used by an author in a given work, citing every passage in which each appears."

Instead of a dedicated soul(s) writing down phrases from some work for a good share of his life, rearranging them, and finally having them put in typeset and published, a computer will read a text from punched cards or magnetic tape and will do the rearranging itself and produce a concordance in a form which may readily be photographed and printed by offset printing, bypassing altogether typeset printing and the expense in setting up the type for such a large volume.

Concordances are not limited to the Bible, but have been produced for many diverse authors and works.[13] As a useful piece of research material including use in word studies, the concordance has a long history, but one which until recently was marked by great effort and long years in compiling. Using a model set up by Paul Tasman from his work with Busa the procedure for concordance preparation using automated techniques involves the reduction of the text into thought units (logical paragraphs), further textual reduction into phrases which are of a suitable size for machine processing (generally some less than 80 characters including spaces), reduction of these phrases into words, indication of the reference, placement and value of the individual words, classification by family, alphabetizing, and indexing of individual words, and the physical association of the individual words "with the text in all places where they appear, prepared in such form that these associations may be useful to researchers

[13] Busa, Roberto, Sancti Thomae Aquinatis Hymnorum Ritualium Varia Specimina Concordantiarum, pp. 12-16.

"in scholarly and statistical studies."[14] This includes what the machine does, taking each word, classifying it by alphabet or by other consideration (e.g., inflection), and then associating its phrase back with it to print it out in a useful form. The place of the scholar and the clerk in this procedure is the marking of text by the scholar, punching the text on punch cards by the clerk, and verifying the cards for accuracy. The machine will take the phrase cards and produce the word cards by itself.[15]

This is essentially the same procedure followed in other mechanical concordance productions. There is varied equipment in use. (Busa started with punch cards and a sorter and moved on to a totally automated system of computer and magnetic tape.) However, most of the work is now done on the computer rather than sorting punch cards for its greater speed and flexibility. The computer was in full use by 1957 when Tasman and Busa worked out the indexing for the Dead Sea Scrolls on an IBM 705 computer, John W. Ellison had published the Nelson's Complete Concordance from Remington Rand's Univac I, and Cornell University launched its computer produced concordance series.[16] This was the same year that Guy Montgomery's Concordance to the Poetical Works of Dryden was

---

[14] Tasman, op. cit., p. 253.

[15] Ibid., p. 254.

[16] Fogel, E. G., "Electronic Computers and Elizabethan Texts", in Studies in Bibliography 15:16-17, 1962.

published by the University of California Press from 240,000 manually indexed cards which were checked and published electronically.

To date Busa's work would hold the record. His work with the writings of St. Thomas Aquinas, and now some of Thomas's sources, totaled 15 million words, 2.5 million lines, in 8 languages and 3 different alphabets as of late 1964.[17] The other sizable effort to date has been the Cornell series. At present, concordances for Matthew Arnold (1959) and William Butler Yeats (1963) have been published by the Cornell University Press under the editorship of Stephen M. Parrish, as has one for Emily Dickenson under Stanford Rosenbaum (1964). The concordance for William Blake under David Erdman is in the post computer phases and should be available shortly.[18] As mentioned above, the Busa center in Italy is also working on indexing the Dead Sea Scrolls, and several other concordance and indexing uses of the computer are being attempted in other places throughout the world, including, for example, several concordances of "The Mahabharata" being done at the American Institute of Indian Studies at Deccan College in India.[19]

---

[17] Busa, Roberto, "An Inventory of Fifteen Million Words", in Proceedings, Literary Data Processing Conference, Yorktown Heights, N.Y., 1964, p. 66.

[18] Painter, J. A., "Implications of the Cornell Concordances for Computing", in Proceedings, Literary Data Processing Conference, Yorktown Heights, N.Y., 1964, p. 160.

[19] Bowles, E. A., "Computerized Research in the Humanities: A Survey", in ACLS Newsletter 16:20, May, 1965.

There are several problems encountered in the programming and running of computer-produced concordances. An initial problem is the problem of input which will be discussed below. Beyond that problem is the consideration as to what texts are going to be used, old spelling vs new spelling (where appropriate), what to do about textual variants, and other textual considerations.[20] After the text is in the machine the homograph and paradigm problems arise as they did with automatic translation. They are met with much the same solutions although these are not such pressing problems for concordances, for the reader can see from the text which accompanies each word how it is used and decide the meaning for himself. One of the procedures evolved by Busa to deal more adequately with inflections is called 'lemmatizing'. In lemmatizing, a scholar goes through and assigns to each inflection form a lemma, or title, under which all the inflections of a single root word will be grouped.[21] The problem posed by hyphenated words (if the hyphen is a letter, the second word is not catalogued; if it is a space, then the whole hyphenated form is not catalogued) is removed by programming the computer to treat the form as one word, but to go to a subroutine which cross-references the second

---

[20] Fogel, "Electronic Computers and Elizabethan Texts", op. cit., pp. 22-23.

[21] Busa, "An Inventory of Fifteen Million Words", op. cit., p. 70.

word.[22] The problem of output also arises in dealing with concordances. In Busa's concordance there would be 500 volumes of 500 pages each (15 million lines) if all the words were printed out. Out of the 2 million words he notes 1900 which occur more than 100 times.[23] It is unfeasable to print a complete concordance, and some standard has to be set up by which the most common (and usually the least distinctive or meaningful) words can be excluded. Ellison, as we shall see, was able to use previous work to help in the decision, depending on information from Strong's concordance of the Authorized Version of the Bible to help the machine decide in borderline cases (e.g., 'has' as a possessive verb, and as an auxiliary verb). Parrish gave the computer a list of 150 words that it should not put out.[24]

Finally, the format may be problematical. It would be ideal if the computer were to print out with each word the thought phrase in which it was found. However, the computer cannot sense meaning, so other ways have to be devised. Busa fed the text in in the appropriate phrases, but the common Key Word In Context concordance program from IBM prints out a certain number of characters to the right and to the

---

[22]Parrish, S. M., "Problems in the Making of Computer Concordances", in Studies in Bibliography 15:7, 1962

[23]Busa, "An Inventory of Fifteen Million Words", op. cit., p. 77.

[24]Parrish, op. cit., p. 4.

left of the word in question, enabling data to be fed in without 'prephrasing'.[25] Either of these ways is, of course, adequate, but the latter comes from a need for easier text input (discussed below) in which less human editing is necessary. The former allows the context print out to be related more closely to the meaning phrase. At this point in computer concordances it is a case of having or eating one's cake. Either the extra editing is done, or the phrasing around the word is arbitrary.[26]

Before moving into the vast area of stylistic analysis with the computer, a final word about a non-linguistic function of the computer in literary studies. This is the area of textual criticism and editing. In his Methods of Textual Editing Vinton A. Dearing announced the completion of a new program to record variant readings using the IBM 7090 computer.[27] (This was five years after the completion of Ellison's thesis on the same topic which is discussed in the next chapter.) He outlined the use of the computer for textual studies in the area of collation of variant texts. Thus

---

[25]Smith, R. H., Jr., "A Computer Program to Generate a Text Concordance", in Proceedings, Literary Data Processing Conference, Yorktown Heights, N.Y., 1964, p. 114.

[26]An interesting look at the spiritual side of concordance making is provided by Fr. Busa: "On my part, my final advice to anyone who would like to make inventories of millions of words, is to take it as a marvellous way to expiate his own personal sins!" from "An Inventory of Fifteen Million Words", op. cit., p. 78.

[27]p. 1, 1962.

the prospective editor of a critical edition of some work(s) would have an automated assistant helping prepare the text. This does not relieve the responsibility of the editor in choosing the text which is the starting point, but it does automate the cataloguing of the text, helps in consideration of textual archetype determination, and helps in consideration of the advisability of emending the presumed archetype in view of authorial change from the testimony of other authoritative manuscripts.

Linguistic Literary Analysis

"Words together form a pattern of sounds and associated sounds, ideas and associated ideas, and the tendency to use certain patterns is the style of the author."[28] Analysis of this style is a large part of what is called 'literary analysis'. More accurately, the literary analysis is the study of the components of style, the individual factors which go to make up style. In the past such study has generally been limited to notations of frequency of particular words, peculiar expression, grammatical constructions or the like in order to further a theory.[29] However, within the last

---

[28] Vincent, E. R., "Mechanical Aids for the Study of Language and Literary Style", in Literature and Science, International Federation for Modern Languages and Literatures, p. 57.

[29] Wake, W. C., "The Authenticity of the Pauline Epistles", in The Hibbert Journal 47:50, October, 1948.

thirty years the introduction of statistical theory and the use of computers to the field of literary analysis has effected a remarkable change. The means of analysis are assuming more statistical sophistication in their processes and in the analysis of their results, and through the use of computers they are able to include greater quantities of data as well as more minute detail in its handling and analysis.

The impetus for linguistic literary analysis has come from two general sources. First is the lexical computer fields which are looking for more precise definition of the syntactical structures within which to perform their function, e.g., mechanical translation. The other chief source of the pursuit of literary analysis has come from studies in authorship determination. In regards to New Testament study this will be more fully discussed in the next chapter. Suffice it to say here that this has also been a concern of both classicists and historians of many different areas. The purpose of the Mosteller and Wallace work, while more centered in illustrating statistical theory, is also to try to solve an older problem in authorship attribution.[30] The setting for most stylistic and literary studies in recent years has been the problem of authorship determination, for it is this analysis which, as we shall see, lends itself most easily to the application and experimentation of statistics and use of the computer in stylistic and linguistic theory.

---

[30]Mosteller, F., and Wallace, D. L., Inference and Disputed Authorship: The Federalist, p. 1.

In the study of literary style there are two major presuppositions which underly the conclusive use of the data. One is the assumption that the writings of an individual reflect his mind, and that in some sense this and the resultant stylistic habits are distinct and somewhat unique to him. The other undergirding assumption is that these stylistic habits "persist for long periods and over a variety of subject matters, so that it is possible to establish statistical indicators of authorship, that is to say numerically expressible stylistic habits ... ".[31]

Armed with assumptions that each author is somewhat unique in his style, that this style is consistent throughout his writings, and that it is quantifiable and, therefore, subject to statistical analysis, the authorship attribution study is ready to be undertaken. The first problem in authorship attribution is what exactly is to be studied -- what particular problem is to be investigated, using what kind of data. The construction of a sample is the first item of procedure. Then the problem of criteria is encountered. (While the more important aspects of the studies done to date will be brought out in what follows, discussion of some of some of the individual studies of particular relevance will be undertaken in the last section of this chapter.)

---

[31] Morton, A. Q., and Levison, M., "Some Indicators of Authorship in Greek Prose", in The Computer and Literary Style, edited by Jacob Leed, p. 154.

In attempting a statistical analysis of the elements of style the construction of a sample is necessary. The sample, which is part of the total population (real or hypothetical total of data applicable), to be subjected to the analysis should meet the qualifications of adequate representation of the total population and sufficient size to produce significant results.[32] The former standard is a priori; unless it is checked against other samples from the same population its represntativity must be determined by other means. The latter standard is a posteriori. It is not known whether a sample is of sufficient size to yield meaningful results until the results from the sample in question are known. To be of sufficient size the standard error attributable to the particularity of the samples in the total population must be less than a potentially meaningful disparity found between or among the samples. With the use of computers the necessity of small samples is somewhat lessened. This contributes both by way of more accurate representation through larger samples of the total population and of greater discrimination possible among elements in the analysis.

In statistical literary analysis, the possibilities

---

[32] A sample smaller than the total population itself would be used when it is impractical to deal with the mass of the total population. See Yule's comments in The Statistical Study of Literary Vocabulary, pp. 35 ff.

of random, block, and spread samples are possible. However, Wake rejects the use of random sampling both for the lack of an independent ~~2nd.~~ second random sample within the individual writing and because " authors do not use sentences randomly".[33] He establishes the preferability of block sampling (a single block as the sample) or of spread sampling (several blocks uniformly taken from throughout the whole writing). With the use of the computer it would be possible to forego this concern and utilize all of the data. However, this was not open to the earlier works in literary statistics such as Wake, Yule and Williams.[34]

The nature of the contents of the samples is no small controversy. It is bound up with the question of what criteria are applicable and gain meaningful results. The need for samples of different materials, the works of different authors, is obvious. (This gets to be a bit of a problem in dealing with Greek literature. See below.) However, is the comparison to be done along the lines set up by Mosteller and Wallace by which criteria are established

---

[33] Wake, W.C., "Sentence-Length Distributions of Greek Authors" in Journal of the Royal Statistical Society, Series A, 120:337, 1957.

[34] Wake, opera cit; Yule, G.U., "On Sentence-Length as a Statistical Characteristic of Style in Prose: With Application to Two Cases of Disputed Authorship" in Biometrika 30: 363-390, Jan '39 (hereinafter refered to as "On Sentence-Length..."; and The Statistical Study of Literary Vocabulary, Cambridge, Cambridge University Press, 1944; and Williams, C.B., "A Note on the Statistical Analysis of Sentence-Length as a Criterion of Literary Style" in Biometrika 31: 356-361, March, 1940; and "Studies in the History of Probability and Statistics IV. A Note on Early Statistical Study of Literary Style" in Biometrika 43:248-256, Dec. 1956.

from the comparison of two known authors[35] or by the methodology of Ellegaard setting the sample of one author against the sample of the rest of contemporary literature of that particular genre.[36] Perhaps this latter may be more accurate in picturing a single style against the general style of an age, but the former sample method of two equally discrete entities being statistically contrasted lends itself to much greater detail in setting up statistical means of comparing a third sample to be associated with one or the other of the original pair. This is a sampling optimum which cannot always be met. The lack of a third sample of valid use to the study of two works by a supposed author, one known and one disputed, throws the statistical results in shadow of doubt. The level which the sampling process must meet or seek to achieve in the statistical study of authorship attribution is that of a complete sample of the author's known work of the type in dispute, and a complete sample of the "opponent's" work. Should there be no known contender for authorship, the provision of such a genre sample as Ellegaard[37] is preferable to the mere statistical study of the disputed work and the known works of the attributed author.

The necessary length of the sample for valid results

---

[35]Mosteller, F., and Wallace, D.L., "Notes on an Authorship Problem" in Proceedings of a Harvard Symposium on Digital Computers and their Applications p. 167.

[36]Ellegaard, A., A Statistical Method for Determining Authorship pp. 20-21.

[37]ibid p. 20-21.

is dependent on the nature of the tests and the nature of the means carrying out the tests. Although the possibilities of longer extant English works are much greater, the Greek works able to be examined are usually no more than 500 sentences.[38] It is possible to obtain meaningful results within this limitation as have Morton and Levison[39] and Yule using as low as 120 sentence samples for his sentence-length statistics.[40] Yule's 'K characteristic' study needed text as long as 10,000 words for meaningful results, however,[41]

Textual accuracy and clarity also play a part in the determination of the sample to be used in statistical analysis. This is most problematic with Greek and other classical authors and works. Wake was able to handle this problem by omitting only 2% of the sentences that he considered ambiguous in the placement of the end and by the use of samples that avoided the more corrupt parts of the texts, the beginning and the end.[42]

Thus, adequate representation by the sample and sufficient length for meaningful results play a determinative

---

[38] Wake, "Sentence-Length Distribution of Greek Authors" p. 337.

[39] Morton and Levison, op. cit. e.g. p. 153.

[40] Wake, "Sentence-Length Distribution of Greek Authors", p. 334.

[41] Yule, The Statistical Study of Literary Vocabulary p. 281.

[42] Wake, "Sentence-Length Distribution of Greek Authors", pp. 334-335, 337.

part in the establishment of samples for literary analysis. Within this framework further consideration must be given to the adequacy of the samples qualitatively as well as quantitatively to produce meaningful results, demanding a relationship among the samples that the results of study of that relationship will accurately portray. Due consideration must also be given not only that the sample represent the text, but that it represent the author (who is really what is being studied) through the elimination of corrupted textual passages and the elimination from statistical consideration of those portions of the sample which are ambiguous in terms of the characteristics to be examined.

After what is to be studied has been decided, the question of criteria of study arises. The determination of what elements of style are to be studied, much less which ones are to be used in finally determining (or trying to determine) authorship, is a subject of serious debate. What the criteria are and how they should be arrived at is far from settled.

One of the basic principles of stylistic characteristics is that they be dependent in no way on context, and be consistent throught the works of a particular writer.[43]

---

[43] Milic, L.T., "Unconscious ordering in the Prose of Swift" in Leed, J., ed., The Computer and Literary Style p. 83, and Somers, H.H., "Statistical Methods of Literary Analysis" in Leed, J., supra p. 129.

Unless the characteristics are not tied to the context in which they appear, they cannot be treated as in any significant way indicative of the author's style as opposed to another author's style. This is particularly applicable when the samples being compared are not on the same subject nor, especially, of the same genre. With proportional greater caution contextual elements could be examined if the contexts for the examined samples were very similar. Thus, the use of "federal" or "national" would possibly be significant between two works if both were on the nature of the Presidency, or some such subject. It would in no wise be valid to discriminate between a work on the U. S. tax system and a work on Chinese opium smuggling. It is recognized that within a genre a writer is going to have greater simularity among his works than will exist between inter-genre works. The object in establishing the criteria for statistical study of style is to get the maximum use out of as many criteria as possible with as high a level of detatchment from and independence of the particular context of the individual writing.

There have been several possible criteria tried and used. Among the more generally accepted ones are sentence length and vocabulary studies. Others have been suggested such as the list of L. Brandwood including syntax, rhythm patterns, clause proportions among type, order, length, and

construction, and parts of speech.[44] Bernard O'Donnel in trying to determine the authorship proportions of the chapters of The O'Ruddy by Stephen Crane used 18 different variables, including, among others, words, various types of clauses, major parts of speech, verbals, semi-colons and dashes, metaphor, initial conjunctions and dialogue.[45]

However, the main studies to date have centered around sentence length and vocabulary usage. The work of G. Udney Yule, C.B. Williams, and W.C. Wake[46] has concentrated on the statistics of sentence length variation as the determinative factor in authorship problems. The statistical approach has become increasingly sophisticated since Yule's first article in 1938, but the stylistic criterion remains the similar feature of these works.

However, Yule has also added another factor to his analyses. Using the characteristic 'K' he goes through and checks up on the De Imitatione Christi which he had earlier worked on with sentence-length statistics. The characteristic 'K' is a function of vocabulary use, and it is derived by finding the frequency of noun usage throughout the sample. Besides the numbers of nouns occuring once, twice,

---

[44] Brandwood, L., "Analysing Plato's Style with an Electronic Computer", in Bulletin of the University of London, Institute of Classical Studies, No. 3, 1956, pp. 47-53.

[45] "Stephen Crane's the O'Ruddy: A Problem with Authorship Determination" in Leed, J., ed., op. cit. pp. 109-111.

[46] cf: Note 34, supra, p. .

etc., the actual nouns used interested Yule very much.[47] He found that 'K' was a viable criterion for authorship in some cases and suggested that its application to other parts of speech might enlarge its usefulness.[48]

The work of Mosteller and Wallace, also described by Ivor Francis,[49] revolves around the concept of "key words" or "marker words". Through comparisons of known writings about subjects similar to the ones needing an authorship decision, a list of words, the usage of which separates one from the other, is compiled. The writing in question is then compared with the list and assigned one or the other as author on the basis of its vocabulary association. This presupposes that it will go one way or the other, not both or neither. Mosteller and Wallace used this tack after having tried sentence-length tests and seeing the results for Hamilton and Madison's known writings turn up as identical, 34.55 and 34.59 respectively.[50] Ellegaard also uses a vocabulary study after failing to find sufficient discriminating power in previous studies on sentence-length and Yule's 'K' factor.[51] This study works from the word frequency

---

[47] Yule, G.U., The Statistical Study of Literary Vocabulary, pp. 2-4.

[48] ibid p. 281 ff.

[49] Mosteller and Wallace, opera cit; and Francis, I., "Authorship: An Exposition of a Statistical Approach to the Federalist Dispute" in Leed, J., ed., op. cit., pp. 38-78.

[50] Mosteller and Wallace, "Notes on an Authorship Problem" in Proceedings etc. supra, p. 164.

[51] Ellegaard, op. cit. pp. 10-11.

statistics of vocabulary. He proceeds on the hypothesis that "the relative frequency of a particular text will not be significantly different from its frequency in any other text by the same author." To do this requires great amounts of data including samples of at least 100,000 words for validity of words occuring about 10 times.[52] Ellegaard takes pains to guard against variation from smaller samples necessitated by his data.[53] Even so, the perils of this should be fairly apparent, and one might share John Ellison's interest in using the same proceedure to show from a letter by Thomas Jefferson to his wife in June 1776 that either he did not write the Declaration of Independence, or someone else was having an affair with his wife and signing his name.[54]

Ellegaard and Mosteller and Wallace differ in basic premise from Yule on the significant part of vocabulary. Mosteller and Wallace have based their study on selected known differential words. Ellegaard has done much the same thing using word usage which distinguishes a writer from the general trend of writing. Yule, on the other hand, uses the whole vocabulary use of an author, weighing not his word usage individually and selectively, but in statistical

---

[52] Ellegaard, op. cit., pp. 12-14.
[53] ibid., pp. 15-19.
[54] "Computers and the Testament" in Bowles, Computers in Humanities Research: Readings and Perspectives, p. 166; Ellison is using this against the work of Morton.

consideration of the whole noun (and expandible to other parts of speech) usage. It would seem that Yule's characteristic being a function of language use rather than individual word use would be less open to contextual variation and thus, in general, more dependable.

Semantic Literary Analysis

In the previous discussions of this chapter the lexical and linguistic functions of computers have been discussed. In the former the machine dealt clerically with the words, but in the latter used the words to get at patterns behind the words -- the author's style. Sedelow and Sedelow note two divisions of consideration of style: form and texture.[55] Under form is catalogued the stylistic traits discussed in the section on linguistic analysis as well as the genre of the writing; eg., an abstract, a political tract, etc. Under texture is catalogued the tone or generality of semantic impact, and the patterns of word association, from the interrelated similies and metaphors to the etymologically interrelated content words. The studies discussed above and the ones to be discussed in chapter 3 have singularly avoided dealing with the texture of style.

This has, however, been considered in the work of

---

[55] Sedelow, S.Y. and Sedelow, W.A. "A Preface to Computational Stylistics" in Leed, J., ed., op. cit. p. 3.

John N. Winburne which has become partially automated since his paper in 1962.[56] In the paper Winburne set forth an argument for semantic analysis of texts. The association of identical words repeated, different inflections of the same word, synonyms, and semantic substitutes (not necessarily even grammatical)[57] is shown to provide a pattern which may also be used in the investigation of an author's style. This semantic structure is a pattern like that of sentence-length and vocabulary usage.

The pattern is determined by the occurrence of the pattern elements called "sensemes". These are discrete cl classes of meaning in which are categorized the meaning words of the discourse as suggested in the previous paragraph. These sensemes may be quantified as are any of the linguistic data and dealt with statistically.

This type of analysis is the obvious end of the trend of abstraction from the pure concern of lexical rearranging to the description and quantification of linguistic patterns seen above. The analysis of texts for the semantic structure is an essential part of the total investigation into the structure of language for use in automatic machine translation, in authorship attribution studies, and in the generalized investigation into the structure of language.

---

[56] "Sentence Sequence in Discourse", in Proceedings of the Ninth International Congress of Linguists (1962), pp. 1094-99.

[57] Ibid., p. 1096.

Input Limitations

"Accurate conversion of the printed text into machine-readable form is a major problem.[58] It is probably the major problem plaguing already existing computer processing of literary data in all phases. Simply put, the computer cannot read. This requires that text to be put in and "computed" must be in a different form from which it appears on the printed page. This leaves it open to human slowness and human error. The procedure is roughly the same for all types of literary data processing. The text is typed on to punch cards with a key-punch (like a typewriter). Then it is either printed out and verified by comparison with the original, or it is re-punched by a second operator on a verifier which locks if a discrepancy from the original is punched. The operator must then check through and make the correction on the original or the 'verifier' punching. The former method is open to error from eyes moving from original text to printed text to original, etc. The second will let an error typed by two different typists pass entirely without question. Sibenz and Devine were not content to let even this level of error creep into their Tertullian concordance. They developed a program to sort out all one- and two- occurrence words "on the assumption that misspellings from accidental

---

[58] Sibenz, J. K., and Devine, J. G., "Concordances to the Works of the Early Christian Writers", in Proceedings, Literary Data Processing Conference, Yorktown Heights, N.Y., 1694, p. 132.

jottings or splitting of words would not repeat themselves more than once or twice."[59] With a tape of 50,000 Latin words from Busa in Milan they eliminated all the words which were found in the dictionary tape and printed out those which were not. "Though this approach may not be infallible by itself, it proved most successful with several other overlapping checks."[60]

The need is either for some means of automatic verification and correction by the computer or for some means whereby the text may be put in directly from the printed page. As to the former, this would not be possible without a similar amount of effort as is required by hand checking. The machine does not know an error unless it is told specifically. To find out would require the same effort whether the machine did the correcting or the person. As to the latter possibility, it is here that the greatest hope for improvement lies. Robert J. Potter reported on the progress IBM is making with a machine that will "read".[61] In the report he cites the progress made with the technique of character recognition but notes that associated problems are now causing delay in

---

[59] Devine, J. G., "Computer-Generated Concordances and Related Techniques in the Study of Theology", in Computers in Humanistic Research: Readings and Perspectives, edited by E. A. Bowles, p. 172

[60] Ibid., p. 172.

[61] Potter, R. J., "On Optical Character Recognition", in Proceedings, Literary Data Processing Conference, Yorktown Heights, N.Y., 1964, pp. 306-323.

perfecting the apparatus.[62] The problem at this point lies just there -- 'almost, but not quite'. The problems are elementary, like how will the machine know where to look on a page for the lines of text, and the different kinds of type face which are used in books, but at this point they are problematical to the completion of a character recognition system -- a 'reading machine'.

The utilization of direct input from source is the only alternative presently open. This involves preparation of the original source of the text in machine-readable form. For most of the work being done with the computer studies of the New Testament text this is not really a 'viable option'.

The problem of input remains. It is a serious bottleneck both in time and expense to fuller use of automatic literary data processing. Its solution would most greatly benefit those studies and programs which use direct text input. For those studies which require editing of text or textual alteration to enable the machine to recognize extratextual distinctions (e.g., homographs) the present method will remain standard for much longer. It is a general rule that the further the study is abstracting in its analysis from the raw text, the harder it will be to utilize direct input of text. A concordance would have little problem handling it, but for the type of analysis indicated by the rhythmic,

---

[62] Ibid., p. 323.

metric, or semantic analyses direct textual input would be practical only when the machine is able to recognize the rhythm, meter, or meaning in the words using only the words themselves.

The lack of a Greek typing and printing element also hampered work in Greek. However, there is now available for some IBM equipment just such an element[63] which will greatly speed up the analysis of Greek through simplifying the procedure of input and output.

Literary Analyses of Greek Texts

This section will discuss the studies which have appeared concerning Greek texts. The studies in literary analysis which use the Greek New Testament have been placed together in Chapter 3. What follows here is intended as background to proceeding into the linguistic analysis of the New Testament Greek text. These studies have, even more than those in other languages, direct bearing on the consideration in the next chapter.

The literary study of Greek in the terms noted in the 'Linguistic Literary Analysis' section of this chapter is marked by several peculiar problems. In undertaking to study statistics of sentence-length variations, ancient Greek presents somewhat of a difficulty in determining exactly what a

---

[63]Personal letter to this writer from the Rev. Walter L. Pragnell, June 21, 1966.

'sentence' is. The Greek of the millenium centering on 100 BC was written in 'periods'. If the sentence end -- the period end -- were not sufficiently clear, perhaps a paragraphos would have been inserted, a short line drawn under the first few letters of the line of writing containing the break. In most cases this is determinable without ambiguity, and the period length is taken as the statistical sentence.[64] The difference of colon and full stop is thus not recognized by Wake in calculating sentence length. As to the modern coincidence with the author's intentions, Wake stated:

> "If obvious interpolation and ambiguous passages are avoided, there is no real reason for supposing that the remaining material, where continuous prose, is not substantially as the author left it, and that the lengths of the periods reflect those in the author's autograph copy."[65]

Wake showed high confidence in being able to accurately assess the author's intention, but he does dismiss the sentences which he counts as ambiguous.[66]

Wake further considered the effect of editing and the resulting changes in punctuation, and he concluded that the

---

[64] Wake, "Sentence-Length Distribution of Greek Authors", op. cit., p. 334.

[65] Ibid., pp. 334-335.

[66] Ibid., p, 335. He excluded 2% as ambiguous. He considered this the upper threshold for study of documents as sufficiently incorrupted for valid evaluation.

difference in punctuation of sentences among editions was insignificant statistically in light of the error built into random sampling.[67]

The other chief difficulty encountered in trying to deal with Greek texts is their length. They are usually no more than 500 sentences.[68] This is not a particularly large sample with which to work. Thus, if textual corruption or some other debilitating factor becomes involved, the adequacy of the sample text to yield valid, meaningful results statistically is radically reduced.

The free use of ascription of name, both by booksellers and by schools of that time[69], also hampers statistical study as the amounts of data in the 'known authorship' category are often minimal. The statistical application to authorship attribution studies is diminished in certainty due to a reduced amount of certain data with which disputed works might be compared.

W. C. Wake in his study of "Sentence-Length Distribution of Greek Authors" which appeared nine years after his first article in 1948[70], followed up the work begun by Yule[71]

---

[67] Ibid., p. 337.

[68] Ibid., p. 337.

[69] Morton and Levison, op. cit., p. 141.

[70] "The Authenticity of the Pauline Epistles".

[71] "On Sentence-Length ...", op. cit.

on sentence-length distribution. He applied this principle with spread sampling to try to determine the authorship of Plato's Seventh Letter and Aristotle's Ethics. He reported favorably on Plato's Letter[72] and concluded from his rather lengthy study that the skew distributions of sentence-lengths in continuous prose are constant enough to be used as "objective criteria of authorship style."[73] Among the samples for any author the intra-sample variances are only those expected in random sampling. The work of Plato departed from this, but Wake attributes this to the particular style and literary form (dialogue) used by Plato.[74]

Morton and Levison adopted Wake's work and expanded it as well as adding more tests of their own.[75] The "Tables" section of Morton and McLeman's Paul, The Man and the Myth also contains many tables concerning such statistics for Greek authors of this time. Morton and Levison concluded from their analysis of forth Greek prose writers of this time that "in every case the differences between works in the same literary form are only those expected in random sampling.[76]

---

[72]Wake, "Sentence-Length Distributions of Greek Authors", op. cit., p. 343.

[73]Ibid., p. 345.

[74]Ibid., p. 345.

[75]Morton and Levison, op. cit., p. 142 ff.

[76]Ibid., p. 142.

Brandwood set out an ambitious program for computer study of Plato's style.[77] From an investigation of sentence structure and clausal word order with the investigator doing the analysis and the machine the compiling and coordinating, Brandwood further suggested machine run word counts, the analysis of rhythm -- long and short syllable pattern, the analysis of syntax and clausal use, and the analysis of parts of speech statistics.[78] An interesting pursuit further along this line has been James T. McDonough, Jr.'s study of the Iliad according to structural metrics and its implications for humanistic research on the computer.[79]

---

[77] Brandwood, op. cit., pp. 45-54.

[78] Ibid., pp. 45-53.

[79] McDonough, J. T., Jr., "Homer, the Humanities, and IBM", in Proceedings, Literary Data Processing Conference, Yorktown Heights, N.Y., 1964, pp. 25-36.

CHAPTER III

THE COMPUTER IN NEW TESTAMENT STUDY

In this chapter is discussed the application of the various procedures and methods seen in the last chapter to the study of the New Testament. The first section will review the use of the computer in dealing with lexical concerns in the New Testament text. The section on the work of W. C. Wake will introduce the use of statistical linguistic analysis to the literary problems of the New Testament. This is taken over and developed by A. Q. Morton and his associates. Using the system he attributed to the ancients[1], he is considered responsible (unless noted) for the work which goes under his name and that of his joint authors since he has been the one person involved with all the work that bears his name. It is Morton who supervises the application of the computer to these problems and so grandly announces its triumphs. It remains to be seen how adequately these results may be attributed to Paul as to Morton. Not to leave the field to Morton alone, the work of H. H. Somers is considered. Although not of the significance or detail of Morton's work, it is an interesting approach to the same problem with some strikingly different answers.

---

[1]Morton, A. Q., and Levison, M., "Some Indicators of Authorship in Greek Prose", in The Computer and Literary Style, edited by J. Leed, p. 141.

Lexical Applications

In the study of the text of the Greek New Testament the application of computer technique by John W. Ellison has been monumental. Faced with the problem of two percent of the extant manuscripts receiving most of the study, and all but twenty percent of the miniscule manuscripts left unstudied, Ellison sought to speed up the process of studying these works which are of such value to New Testament research.[2] Even the study itself was not consistent. Scholars used varying criteria in comparing the manuscripts and did it with no organized pattern -- some studying some points, others working on different points.[3] Thus the need both for increased range of study and for systematic research along universally standard lines appeared to be pressing.

The computer seemed a possible solution for its speed and for its insistence on strictly objective, logical application of tests to data. To utilize the computer Ellison first had to coordinate and construct a precise method for analysing the data of textual variance. He noted eight basic kinds of variant readings: omissions from the standard text, substitution, addition, inversion of word order, proper name

---

[2]Ellison, J. W., The Use of Electronic Computers in the Study of the Greek New Testament Text, hereinafter referred to as The Use of Electronic Computers ... , pp. 4-5.

[3]Ellison, J. W., "Computers and the Testaments", in Computers in Humanistic Research: Readings and Perspectives, edited by E. A. Bowles, p. 162.

spelling, itacism (slide of vowels toward eta), case and tense differences changing the meaning of the sentence, and nonsense spelling errors.[4] A priority system was established for recording the variants from several texts. This controlled the program so that the variances from the standard text will be noted as an omission (if there is one), a substitution (if there is one, and no omission), etc. This was to prevent a multiple listing of variance at a single position.[5] The program was prepared to compare "collations of manuscripts to produce tables listing the number of differences between any pair of manuscripts, according to kinds of variant readings."[6]

In order to prove that this method produced valid conclusions, Ellison then sought to arrange the texts by categories, each category being the group of texts most like one another. This would be checked against the listing of manuscript groupings established by earlier scholars without the use of the computer.[7] In his study, 307 of the 309 manuscripts considered fell into the same groupings from older methods.[8]

---

[4] Ibid., p. 163.

[5] Ibid., p. 163.

[6] Ellison, The Use of Electronic Computers ..., op. cit., p. 5.

[7] Ibid., pp. 70 ff.

[8] Ellison, "Computers and the Testaments", op. cit., p.165.

Working with the sample of 309 manuscripts of the tenth chapter of St. Luke's Gospel (fifteen verses) the program ran through 95,000 pairs of comparisons contrasting with the older method of comparing one manuscript with thirty to forty others.[9] It seems to have borne out the general conclusions regarding manuscript collations reached by the older manual methods,[10] but did so with a great range of manuscripts, greater speed and accuracy, and an objective system of comparison -- all variations, not only the ones which the scholars believe to be significant.[11]

The collation, however, still remained a manual procedure. The time spent was no more for doing the work for the computer than for preparing a collation against a single manuscript, "because it requires only the consultation of a master list of variants and their identifying numbers, and transmitting the list of appropriate numbers."[12]

The other of Ellison's monumental works in mechanical lexical application of the computer is the Nelson's Complete Concordance of the Revised Standard Version Bible which he edited. He supervised the preparation of the text of the

---

[9] Ibid., p. 165.

[10] Ellison, The Use of Electronic Computers ..., op. cit., p. 93.

[11] Ellison, "Computers and the Testaments", op. cit., pp. 162,165

[12] Ellison, The Use of Electronic Computers ..., op. cit., p. 5.

Revised Standard Version of the Bible into a form which the machine could read and also supervised the processing of the text by the Remington Rand Univac.

In dealing with some of the problems noted in concordance preparation in the last chapter, Ellison used the concordance of the Authorized Version prepared by James Strong to cut down on the 'have' and 'will' concordizing.[13] The 'haves' and the 'wills' which were in sentences in which Strong indicated a non-auxiliary function would be found were kept in the RSV concordance whether each actual occurrence was an auxiliary verb or not. There were 132 other words which the machine was told to disregard in compiling the concordance, such words as 'and', 'if', 'of', and 'is', the inclusion of which would have enlarged the 2157 page concordance by two and one-half times.[14] The decision was justified not only on practical grounds, but that there was no text of any significance "made up of these words alone."[15]

The problem of printing the context for the concordance also had to be dealt with. A system was developed to print the context phrases as the words between punctuation marks.[16]

---

[13] Ellison, J. W., ed., Nelson's Complete Concordance of the Revised Standard Version Bible, p. iv.

[14] Ibid., p. vi.

[15] Ibid., p. vi.

[16] Cook, C. M., "Automation Comes to the Bible", in The Christian Century 74:893, July 24, 1957.

This gives a much more consistently informative and helpful context than simply defining the context phrase by a certain number of characters before and after the word in question. (See the discussion of this problem in Chapter 3.) This also allowed the flexibility of not being forced to feed the text into the computer after already having gone through and divided it into phrases, but yet getting a phrase for context more directly related to the text itself.

The Analysis of W. C. Wake

One of the interesting issues raised by modern critical scholarship of the New Testament has been the questioning of the veracity of traditional author assignments of the books of the New Testament. Authorship attribution and the work that goes into its study have become an ever-present part of contemporary New Testament criticism. The analysis of the writings involved have provided a somewhat objective base from which to draw conclusions in this regard, but these analyses have not been consistent, objective, or thorough. They are in a fair measure dependent upon the intuition and the mathematical aptitude of the particular critic. This is further taken up in the first section of Chapter 4. The rigorous consistency and demanding logic of an analysis performed by the computer has much to offer this field of scholarship, as did John Ellison's work have much to offer to the field

of textual study. In establishing an analytical program using the computer to examine the style of the given writers, or supposed writers, the criteria must be objective, independent of judgment in their application, and concrete, in the form or use of the words themselves. (This particular type of analysis does not allow for a discrimination by use of 'themes' of 'convincing style', except as these can be demonstrated from the words themselves.)

One of the earliest attempts to apply sophisticated statistical theory to the problem of authorship determination was that of the statistician William C. Wake in 1948.[17] In that analysis he constructed a statistically viable sample and tested it using the statistics of sentence-length distributions set forth by Yule nearly ten years before.[18] Wake was impressed with this study for its use of the whole work, instead of isolated words, and for its use of valuable statistical calculations.[19]

In constructing the samples with which he would work Wake used the sentences of the Greek text breaking at the colons and the full period stops. He excluded various

---

[17]"The Authenticity of the Pauline Epistles", in The Hibbert Journal 47:50-55, December, 1948.

[18]Yule, G. U., "On Sentence-Length as a Statistical Characteristic of Style in Prose", in Biometrika 30:363-390, January, 1939.

[19]Wake, op. cit., pp. 50-51.

mitigating factors like quotations and the name lists, e.g., at the end of Romans. His text was taken from the Oxford edition of the Textus Receptus of 1863 although he saw the prospective value of using a more modern critical edition, especially in dealing with II Corinthians i-ix.[20] With this structure for the sample to be analyzed, he set up samples from the group of Epistles attributed to St. Paul. These range from 24 to 442 sentences and represent the bulk prose of the author(s).[21] For the purposes of the study, II Corinthians was separated into II Corinthians i-ix and II Corinthians x-xiii (the Severe Letter).

The calculation of sentence-length distributions turned up the results which Wake stated were not in accord with his experience of single author distributions. He hypothesized that this could be from interpolation, correction, or extreme but normal variability.[22] This left the possibility that these were all part of one discrete population,[23] that there were several discrete populations within the total group, or that the samples were really unrelated.

Using Analysis of Variance he broke down the samples and examined them singly to try to establish an internal

---

[20] Ibid., pp. 52,54.

[21] Ibid., pp. 51-52.

[22] Ibid., p. 52.

[23] Used here meaning a group that is statistically homogeneous in the criteria applied.

random error statistic.[24] The results of this convinced Wake that these were not all of a single population. By further comparison of the samples he established that there were two definite populations into which the more important Epistles fell.[25] As for Hebrews, it was "unambiguously excluded from either of them" which came as "no surprise".[26] In the first group:

> "..all methods employed agree in finding that 'I Corinthians', 'II Corinthians' x to xiii (The Severe Letter), 'Galatians', and 'Romans' have sentence-length distributions which are stistically indistinguishable and homogeneous. If any other Epistle is added to this group the constants become statistically heterogeneous."[27]

In the second group he included I Thessalonians, Colossians, Philippians, and probably II Thessalonians although the latter is too short to be decisive.[28] While Ephesians is close to Group II in several characteristics, its shorter sentence characteristics set it apart from Group II. I Corinthians is mutilated and perhaps textually corrupt and should also be omitted, as should the Timothy Epistles, which seem to be related statistically to none of the others, and Philemon and Titus as too short for any reliable results.[29]

---

[24] Wake, op. cit., p. 52.
[25] Ibid., p. 52.
[26] Ibid., p. 52.
[27] Ibid., p. 53. (italics his)
[28] Ibid., p. 53.
[29] Ibid., p. 54.

Wake next examined the internal evidence of the two groups to try to find any common distinguishing factor there. He noted that in the introductions of the Group II Epistles Paul is uniformly linked with Timothy. In Group I the introductions lack mention of Timothy (II Cor x-xiii even lacks an introduction), and where Paul is associated with another person, it is Sosthenes (I Cor).[30] While "the most obvious explanation for the existence of two separate and distinct groups is that they are the work of different authors",[31] Wake also cited evidence from Acts and other internal evidence which casts doubt on the lack of a Pauline connection for the Epistles in Group II. He concluded that the more likely hypotheses are that Group II represents joint letters of Timothy and Paul, that Timothy is the sole author, or that Timothy wrote on Paul's instructions, but not with his phraseology, as an amanuensis.[32] Wake did not attempt to decide definitely among the hypotheses, but noted them as the 'live options' with his methodology. The Pauline authorship for Group I is "fairly well established", and Wake is willing to let that stand.[33]

---

[30] Ibid., p. 54.

[31] Ibid., p. 54.

[32] Ibid., p. 54.

[33] Ibid., p. 54.

Wake's analysis is quite modest in its claims and also quite within the accepted results of earlier scholarship. The statistical model is applied not in support for a theory of Pauline authorship but as an exercise in the applicability of a statistical method to problems in New Testament criticism. Although Wake's work on this analysis was done before the use of a computer was an unexceptional occurrence, the applicability of the computer to this type of analysis is quite clear. It is quite within the capacity of the computer to count words and to do even more sophisticated statistical studies than Wake attempted. And, in addition, the computer will perform the studies more rapidly, with greater accuracy, and with much larger quantities of data.

The Work of A. Q. Morton

In dealing with the work of Andrew Q. Morton there are two considerations which have to be examined before the core of his work which is relevant to this study can be taken up. The first of these is what he writes (jointly). There is running through his work expression of irrelevant (for our purposes) opinion which sometimes takes more than half of the pages (e.g., Christianity in the Computer Age). The third chapter of the first section of that book has some twelve pages on literary analysis -- the computer meat of that work. "The rest of the book could have been written

"by von Harnack."[34] Morton has a definite axe to grind, and grind it he does -- against the Church and organized religion in most of Christianity in the Computer Age, and generally against those who oppose his methodology and conclusions. His remark about the saddening sight of "so many scholars standing naked before a new idea"[35] is exemplary of the attitude taken throughout his works.[36] To get the 'relevant' material for this present investigation the opinions and statements about extraneous matters have to be shelved. In the case of Christianity in the Computer Age that includes most of it. Fortunately, there are greater amounts of material elsewhere.

The second consideration is the development which has taken place in the work of Morton. The death of Macgregor in 1963 marks something of a watershed in the published work of Morton. The two books which he authored jointly with G. H. C. Macgregor[37] were based chiefly on ideas and procedures

---

[34]Rhys, J. H., review, in Anglican Theological Review 48:118, January, 1966.

[35]Morton, A. Q., and Macgregor, G. H. C., The Structure of Luke and Acts, p. 6.

[36]Leicester, Ronald, Review of Paul, The Man and the Myth, in The Churchman (London) 80:324, Winter, 1966, "And like all the books in which A. Q. Morton is joint author the tone and spirit of the writing is deplorable. All is presented in a thoroughly bad-tempered, spoilt-child mood."

[37]The Structure of the Fourth Gospel, and The Structure of Luke and Acts.

developed before the computer was introduced into their work. the analyses of the structural and textual problems of St. John's Gospel, St. Luke's Gospel, and the Acts of the Apostles were centered around the effects of a predetermined amount of papyrus available to the writer and his desire to use every last line.[38] The theory of two sources of the Fourth Gospel endorsed by Macgregor and a slightly altered theory from B. F. Streeter concerning the sources of Luke are shown to accord with the method shown by Morton. His slight changes in Streeter's Proto-Luke hypothesis just happen to even things up for Morton's figuring in the matter, but they do not convince his critics that the changes were textually justified.[39] This type of analysis based on word counts and line counts per probable unit of ancient writing is not the work which has brought Morton the most fame (or notoriety) nor has been his most substantial contribution to the use of the computer in the study of the New Testament.

That came as the result of Dr. A. D. Booth's suggestion that the computer might be used by Morton and Macgregor and their decision to use it to investigate the styles of

---

[38]Macgregor and Morton, op. cit., pp. 40-44, and Morton and Macgregor, The Structure of Luke and Acts, pp. 17-20.

[39]Houlden, J. L., Review of The Structure of Luke and Acts, in The Journal of Theological Studies, new series, 17:141-142, April, 1966.

Greek authors for use in determination of authorship.[40] Published on the same day as The Structure of Luke and Acts,[41] Christianity in the Computer Age offers a short look at things to come in Chapter 3 of the first section.

The work of W. C. Wake in sentence-length distribution statistics was reviewed.[42] The resulting isolation of Romans, I and II Corinthians, and Galatians was noted, and the only criticism offered being that the text used by Wake was not compared with more modern texts.[43] Further, using a set of examples nominated by some Classical scholars, lists were compiled of those works of generally accepted authorship, those of traditionally attributed authorship which is now debunked, and those whose attributed authorship reliability is uncertain. The establishment of what habits of composition which were so ingrained as to be unvarying was the object of the search which would then check the prospective criteria.[44] To be considered valid, the criteria had to accept all the 'good' works, reject all of the 'false' works, and give mixed results about the 'uncertains'. The criteria,

---

[40] Morton and Macgregor, The Structure of Luke and Acts, pp. 5-6.

[41] Dinwoodie, Cameron, review of both books, in The Scottish Journal of Theology 18:204, June, 1965.

[42] Morton and McLeman, Christianity in the Computer Age, p. 26-27.

[43] Ibid., p. 27

[44] Ibid., pp. 28-29

e.g., a frequency count of kai, then had to be checked to determine if variance within individual works was below the variance between works. The variance in a raw kai-per-word count was noticed, with kai being in a greater proportion in those works with less than 100 sentences. This proportion falls off even more as greater numbers of sentences are encountered in Greek prose.[45] In using this test, only samples from works of similar size whould be used for dependable results.[46] Another criterion, the appearance of de at the beginnings of sentences, appeared to be valid, given works of at least 100 sentences and a maximum of 20 years difference in composition (the use of de at the beginnings of sentences seemed to drift slowly over time).[47]

Applied to the Pauline Epistles these tests segregated Romans, I and II Corinthians, and Galatians from the rest. The differences in this group are not so significant as to be associated with different authorship.[48] It is interesting to note that Philemon is added to this group by Morton "for there is nothing in Philemon which makes it unlikely to be by Paul."[49]

Reaction to this volume was quite harsh. Aside from the exceptions taken to the whole tenor of the book (supra),

---

[45] Ibid., p. 31.
[46] Ibid., p. 32.
[47] Ibid., p. 32.
[48] Ibid., p. 33.
[49] Ibid., p. 32.

the criticism of the statistical theory was hard and thorough. John Ellison attacked Morton's use of his statistics and his apparent use of more criteria than he published.[50] Dinwoodie goes after not only the use of the statistics, but the figures themselves.[51] He pointed out several discrepancies between the figures in the text and the figures presented in the table. This kind of attack is fairly devastating to the argument which Morton would try to expound.

The question of a full explanation was met in Paul, The Man and the Myth.[52] The statistical theory and application are outlined at length, and extensive tables are given (54 tables on 79 pages). The stated purpose of the work was to try to establish "the authorship of the Pauline Epistles on an objective basis."[53] With a lack of external evidence offering any conclusive proof, most reliance must be put on stylistic evidence.[54] The use of analysis of the choice of synonymous alternatives as put forth by Ellegaard[55] was rejected due to the limited size of the Pauline corpus, and

---

[50]Ellison, J. W., Review of Christianity in the Computer Age, in The Journal of Biblical Literature 84:190-191, 1965.

[51]Dinwoodie, op. cit., pp. 210-212.

[52]Morton, A. Q., and McLeman, J., Paul, The Man and the Myth.

[53]Ibid., p. 42.

[54]Ibid., p. 43.

[55]See pp. 40-41 supra.

the analysis of common word usage was appropriated.[56] The statistics of sentence-length frequency distributions were applied to several Greek prose authors, and the adherence of the results without exception to the limits set up in the analysis was noted leading to full confidence in the reliability of the test.[57]

Since the connection of the definite article to the nouns and adjectives could lead to the influence of subject matters in its analysis, this analysis of the use of this most frequent common word was dropped in favor of using the frequency of kai as a criterion. However, kai also presents problems by its nonconformity to random occurrence expectations, thus vitiating some of its immediate statistical reliability.[58] Within strict limitations kai may be used, including examination of its occurrences per sentence and the spacing of those occurrences.[59]

The use of the statistics of de at the beginnings of sentences is advised with a need for large samples to increase reliability (e.g., with 200 sentences, a variation of 13% is needed for the notation of significant difference).[60]

---

[56] Morton and McLeman, Paul, The Man and the Myth, p. 44.

[57] Ibid., p. 63

[58] Ibid., p. 70

[59] Ibid., p. 80.

[60] Ibid., p. 83.

These criteria are used to examine the Epistles attributed to St. Paul. The results are the same as Wake's in regard to sentence-length distribution (supra) except that II Corinthians was not split by Morton as it was by Wake.[61] The kai testing confirmed this segregation as did a reconciliation of the de testing (removal of Old Testament quotes from Romans 9:16-13:5).[62] The statistical anomalies in Romans, I and II Corinthians, and Galations are all marked by literary anomalies noticed in earlier literary studies so that the statistical indications of unity of authorship of these Epistles remains unmarred.[63] Morton concluded from this that as regards Pauline authorship for any of the other Epistles, "the number of exceptional circumstances needed to reconcile the evidence and the theory strains the credulity in all but the one case, when the hypothesis is that Paul wrote only the four major Epistles."[64] The criteria which Morton used were not substantially different from those used in other studies. The full use of statistics, increased by the use of a computer, especially as seen in the material of Chapters 5 and 6 of Paul, The Man and the Myth is another addition to the thorough methodology which Morton has

---

[61] Ibid., p. 91.

[62] Ibid., p. 93.

[63] Ibid., p. 93.

[64] Ibid., p. 96.

employed. To attack the conclusions he has asserted it is necessary either to attack the applicability of these criteria to all of Greek prose or to find as easy an explanation why they should work for other Greek prose and yet not apply to the Epistles. The only alternative is to attack the procedure and its results (which has been done, see note 51 supra), or to attack the confidence which they demand, posing as a substitute the probability calculations as Mosteller and Wallace have done.[65]

The Contribution of H. H. Somers

An interesting contrast to the work of Wake and Morton is provided by the analysis of H. H. Somers.[66] Somers took the works of Philo Alexandrinus and the Epistles of Paul and set about to determine whether the differences between them are greater than the differences within each, and whether statistical methods can be used to discriminate between them (or within them). For samples to work with he chose ten works of Philo and ten works attributed to St. Paul. These latter included Ephesians, Hebrews, I and II Corinthians, Galatians, Philippians, Colossians, Romans, I Thessalonians, and texts

---

[65] Inference and Disputed Authorship: The Federalist.

[66] "Statistical Methods in Literary Analysis", in The Computer and Literary Style, edited by J. Leed pp. 128-140.

selected by subject.[67] Using such criteria as verbal classifications and prepositions he first applied the Discriminant of Fisher. He noted no overlapping of scores (λ) between the works of Philo and those of Paul and also a homogeneity within the individual collections.[68] These results were confirmed by the T^2 test of Hotelling. From this Somers concludes that "the interpretation of this result indicates that it should be easy to assign an unknown text to one of both collections [(one of the two)], but much more difficult to discriminate within each collection."[69]

Somers then grouped the Pauline corpus into four groups to try to determine the differences among them. The four groups were: I and II Thessalonians, I Corinthians, and Galatians; II Corinthians and Romans; Ephesians, Colossians, Philippians, and Philemon; and Hebrews, I and II Timothy, and Titus. Tests for the use of kai and the use of the article were then run with significant differences coming out of the Kolmogorov-D-test.[70]

The Pauline letters were then compared with a set of Biblical passages taken from books from Genesis to the Apocalypse with the conclusion being drawn from the results that "Paul's

---

[67] Ibid., p. 131.

[68] Ibid., p. 132.

[69] Ibid., p. 133.

[70] Ibid., p. 134.

"Letters are somewhat more heterogeneous than the works of Philo, but not so much as the collection of biblical texts."[71] This is hardly an outstanding surprise.

A Factor Analysis of the Epistles yielded three general discriminants which could be applied to them: a general vocabulary-level factor, a bipolar qualificative vs dynamic factor (verb and substantive opposition), and mental inhibition against complex sentence subordination. Applying the Chatlos Type-Token-Ratio and Somers' own Θ measurement, the small heterogeneity in Paul's letters was strikingly confirmed.[72] However, Somers explains this homogeneity in terms of vocabulary evolution which he illustrates on the following figure.

Table 11. Values of Θ and Evolution of Vocabulary

Year	Θ
52: 1a Thess and 2a Thess	81
55: 1a Cor and Gal	81-82
57: 2a Cor and Rom	82
First Captivity	
60: Eph, Dol, Phil, Phem	82-83
Second Captivity	
68: Hebr, 1a Tim, Ti, 2a Tim	84-87

'a' is a statistica notation having to do with the Type-Token-Ratio and is not a concern here.

Figure 1. Somers' Vocabulary Evolution Scale[73]

[71] Ibid., p. 135.

[72] Ibid., p. 139.

[73] Ibid., p. 139.

Thus the most challenged Epistle, Hebrews, is placed within the evolutionary pattern of Pauline vocabulary. While in no sense has the computer proved that these are all truly Pauline, it has certainly produced statistics which allow it (from the same data which were used to prove non-Pauline authorship, only different testing criteria).

In conclusion Somers cites a large variation in preposition use among the Epistles. Since this may be dependent on the usage of "in Christo" so greatly in Ephesians and Colossians, he cautions that this analysis could be readily influenced by variation in the author's ideas and attitudes.[74] This would result in the same phenomenon observed in the inordinate proportion of seven letter words in Dickens' A Christmas Carol.[75]

---

[74] Ibid., p. 139.

[75] Williams, C. B., "Studies in the History of Probability and Statistics IV. A Note on an Early Statistical Study of Literary Style", in Biometrika 48:255, December, 1956. c.f. "Scrooge".

CHAPTER IV

FURTHER APPLICATION OF THE COMPUTER

The use of the computer in New Testament studies, as in literary studies in general, is of little value if it is not expanded into areas where its capabilities are needed, as well as useful. The logical and mathematical rigor enforced by the computer would be very useful in reevaluating older statistical linguistic studies of the New Testament. There is also a diversity of approach to the application of the computer in literary studies which should be integrated.

If the use of the computer is going to be as great as it can be, it should be used in a procedure which is integrated in its total approach to the problem of literary data processing.

Reconstruction and Validation of Older Studies

One of the contributions expanded use of the computer can make to New Testament research is in the reconsideration of older linguistic studies. The demands of the computer for strict, objective criteria for analysis as well as its ability to perform more extensive and intensive analysis has much to offer to these older efforts.

An example of this in Johannine studies is the work of C.H. Dodd[1] in analyzing the differences between the First

---

[1] The First Epistle of John and the Fourth Gospel.

Epistle and the Gospel of John. He recounted the grammatical words and particles that are used in the Epistle and in the Gospel, compound verbs, idioms and rhetorical figures, and vocabulary differences.[2] These criteria could well afford to be re-examined in light of the latest work on the use of vocabulary as a factor in separating works of different authors. The whole work could be mechanized, allowing it to be applied to other problems and run on other data to check the accuracy and the validity of the results. While at present the parts on sentence structure would require coded input, with moderately sophisticated programming, the rest should be able to be checked by computer. The validity of the criteria in separating other works of different authorship would bear directly on its valid application in the use Dodd made of it.

Likewise, in his work on Ephesians[3] C.L. Mitton made use of comparisons of style and usage in discussing the relationship of Ephesians to the acknowledged Pauline Epistles. He sets up a statistical model for comparing the parallels of Ephesians and the rest of the Pauline Epistles with those of Philippians and the other Epistles.[4] The differences in length of the Epistles are taken into account as is the

---

[2] Ibid., pp. 5-15.

[3] The Epistle to the Ephesians.

[4] Ibid., pp. 107-109.

difference in potential ground for parallels. (Ephesians is compared with and without Colossians.) Also the distribution of the placement of these parallels throughout the Epistles was checked, and the quality of vivid impression of the parallel on the reader was also taken into account.[5] While this latter, more subjective, test could not be automated at this point, the other studies are quite amenable to computer application. The working out of complex statistical problems as well as the rapid location and identification of parallels or identical forms are very much within the capabilities of computers today.

In discussing the authorship of the Pauline Epistles, P.N. Harrison[6] set out several linguistic tests which were intended to show the relationship of the Pastoral Epistles to those generally recognized as of genuine Pauline authorship. By comparing the vocabularies of the Pastorals, the Paulines, and early Second-Century writers, Harrison concluded that the Pastorals were not written by Paul but were the product of a Paulinist who was more nearly in touch with the Second Century than the First.[7] With a computer these figures could be taken from a much wider range of contemporary texts and the hapax logomena could be compared against a

---

[5] Ibid., pp. 114-117.

[6] The Problem of the Pastoral Epistles.

[7] Ibid., pp. 84-86.

fuller set of data. The probabilities for their occurrence as a function of the context also reveals a need for more specific investigation than Harrison has done. A larger, more complete study of the total vocabulary of these Epistles would do much to lend greater credibility to Harrison's assertions.

Each of these studies could benefit from a much more expansive use of the computer in doing the literary analysis which it has recently been given in the total field. The availability of high speed machinery with many workable functions of analysis demands that these total methods be applied in examining problems using the language of the New Testament and analysis of author style.

Further Application of Present Techniques

The use of computers in the translation of the Bible is largely an unexplored area. The applicability and indeed the integration of linguistics and Bible translation is an accomplished fact,[8] but the insights of linguistics need to be applied still further.[9] It is clear that the

---

[8]Smalley, W.A., "The Place of Linguistics in Bible Translation," in The Bible Translator 16:105-112, July, 1965 and Gleason, H.A., "Linguistics in the Service of the Church," in The Hartford Quarterly 1:7-27, 1961.

9. Gleason, op. cit., p. 11.

computer should also be useful to this work and can be of great value. The linguistic and other structural insights made available through computer work for the benefit of mechanical translation are equally useful to the translation by humans. The construction of word lists and frequency counts discussed by Dennett[10] and by Robinson[11] could easily be done by the computer, leaving the researcher to do more work in the field and in application of the results. In providing computer procedures for the use of missionaries in translating the Scriptures the initial problem in this application is the relative unavailability of computers to workers in the field. It is hard to conceive of missionaries in the jungle of South America carrying large computers to do this work. It is neither practical (no electricity usually) nor expedient (for five minutes operation in a month, especially when the availability of servicing may also be a problem) for computers to be placed in the field. However, data processing centers with operational and experimental facilities would be of decided advantage.

---

[10] Dennett, Herbert, "Word-lists in English, Problems of Construction," in The Bible Translator 14:81-87, April, 1963.

[11] Robinson, D.F., "Native Texts and Frequency Counts as Aids to the Translator," in The Bible Translator 14:63-71, April, 1963.

In the linguistic analysis of the Greek New Testament, especially for determination of authorship, a thorough reconstruction of the present procedures is needed. Each writer, whether as Analyst or Commentator[12] on the questions involved, sets up his own criteria and methodology in getting his results. What is now needed is a total collation of the studies to date with the statistics redone and the methods retraced. Much more detailed analysis like the work of H.H. Somers (supra) with Factor Analysis needs to be undertaken. A total effort to recount and reconsider all avenues of statistical and linguistic approach will serve to increase greatly both the significance and the definiteness of the statistical results and the conclusions drawn from them. This is in accord with the call by E.G. Fogel for the mathematical "dressing up" of empirical quantitative studies.[13]

There are many statistical and mathematical procedures and methods which can add greatly to the sophistication of the procedures now used in literary data processing. In a real sense the whole field has only begun to be explored. The work already done may seem like much in retrospect, but

---

[12]Macgregor, G.H.C., and Morton, A.Q., The Structure of the Fourth Gospel, p. ii. (This is not meant here as an exclusive reference, but as indicative of attitude toward the material.)

[13]Fogel, E.G., "The Humanist and The Computer: Vision and Activity," in Proceedings, Literary Data Processing Conference, Yorktown Heights, N.Y., 1964, p. 17.

it will serve only as a point of departure for more detailed, more accurate, and faster methods in dealing with literature in general, and with the New Testament in particular. The computer has in large part provided the capabilities for this expansion, but its impetus, direction, and future guidance have come, and must come, from men. It is up to men to do the judgment of conclusion involved in any analysis. They can do so only when recognizing that their tools, of which the computer is one, are just that -- aids in the more creative and profitable task of inquiring scholarship free from the drudgery which previously took so much time and effort.

The computer is a very great tool of New Testament study. Its utility is enhanced by the proper recognition of its place as a tool among others. Its uses and applications are limited only by man's imagination (and the limits of mathematical action). Moreover, it is a multidiscipline tool, uniting various academic concerns for the benefit of all. The computer is fully applicable to New Testament study in that it is applicable in many disciplines useful to New Testament research and interpretation. The integration of disciplines is a requirement not only for the full utilization of the capacities of computers, but also for the search for

knowledge that is found in the totality of academic inquiry. This is the present state of work with the computer. Its usefulness has only begun to be tapped. Where it leads, and how far, is the province of Man, his abilities and his limitations.

BIBLIOGRAPHY

BIBLIOGRAPHY

Anderson, David L., "Placental Roots", a review of The Secular Promise by Martin Jarrett-Kerr, C.R., The Relevance of Science by C. F. von Weizäcker, and Christianity in the Computer Age by A. Q. Morton and James McLeman, in The Christian Century 82:1290, October 20, 1965.

Bessinger, Jess B., "Computer Techniques for an Old English Concordance", in American Documentation 12:227-229, July, 1961.

Booth, Andrew D.; Brandwood, L.; and Cleave, J. P., Mechanical Resolution of Linguistic Problems, New York, Academic Press, 1958.

Bowles, Edmund A., "Computer Research in the Humanities: A Survey", in ACLS Newsletter 16:7-31, May, 1965.

"The Humanities and the Computer: Some Current Research Problems", in Computers and Automation 15:24-27, April, 1966.

Brandwood, L., "Analysing Plato's Style with an Electronic Computer", in Bulletin of the University of London, Institute of Classical Studies, no. 3, pp. 45-54, 1956.

Bruce, F. F., Review of The Structure of the Fourth Gospel by G. H. C. Macgregor and A. Q. Morton, in The Evangelical Quarterly 34:112-114, April-June, 1962.

Busa, Roberto, S.J., Sancti Thomae Aquitatis Hymnorum Ritualium Varia Specimina Concordantiarum, intro. in Italian and English, Milan, Bocca, 1951.

Cadbury, H. J., Review of The Structure of Luke and Acts by A. Q. Morton and G. H. C. Macgregor, in The Journal of Biblical Literature 84:337-338, 1965.

Cook, Clair M., "Automation Comes to the Bible", in The Christian Century 74:892-894, July 24, 1957.

Dearing, Vinton A., Methods of Textual Editing, Los Angeles, William Andrews Clark Memorial Library, UCLA, 1962.

Delavenay, Emile, An Introduction to Machine Translation, London, Thames and Hudson, 1960.

Dennett, Herbert, "Word-Lists in English, Problems of Construction", in The Bible Translator 14:81-87, April, 1963.

B.6
3

Do
Over

De Tollenaere, F., Nieuwe Wegen in de Lexicologie, with summary in English, Amsterdam, North-Holland Publishing Co., 1963.

Devine, Joseph G., S.J., "Computer-Generated Concordances and Related Techniques in the Study of Theology" in Computers in Humanistic Research: Readings and Perspectives, pp. 170-178, edited by Edmund A. Bowles, New York, Prentice Hall, to be published in late April, 1967.

Dinwoodie, Cameron, Review of Christianity and the Computer by A. Q. Morton and James McLeman and The Structure of Luke and Acts by A. Q. Morton and G. H. C. Macgregor, in The Scottish Journal of Theology 18:204-218, June, 1965.

Dodd, Charles Harold, The First Epistle of John and the Fourth Gospel, reprinted from the Bulletin of the John Rylands Library, vol. 21, no. 1, April, 1937, Manchester, Manchester University Press, 1937.

Ellegaard, Alvar, A Statistical Method for Determining Authorship, Göteborg, The University, 1962.

Ellis, E. Earle, Paul and His Recent Interpreters, Grand Rapids, William B. Eerdmans, 1961.

Ellison, John W., The Use of Electronic Computers in the Study of the Greek New Testament Text, a thesis presented to the Harvard Divinity School in partial fulfillment of the requirements for the degree of Doctor of Theology in New Testament, Cambridge, Mass., April, 1957.

Review of Christianity in the Computer Age by A. Q. Morton and James McLeman, in The Journal of Biblical Literature 84:190-191, 1965.

"Computers and the Testaments", in Computers in Humanistic Research: Readings and Perspectives, pp. 160-169, edited by Edmund A. Bowles, New York, Prentice Hall, to be published in late April, 1967.

editor, Nelson's Complete Concordance of the Revised Standard Version Bible, New York, Thomas Nelson and Sons, 1957.

Fogel, Ephim G., "Electronic Computers and Elizabethan Texts", in Studies in Bibliography: Papers of the Bibliographic Society of the University of Virginia 15:15-31, 1962.

Gleason, Henry Allan, Jr., "Linguistics in the Service of the Church", in The Hartford Quarterly 1:7-27, 1961.

Goetchius, Eugene Van Ness, The Language of the New Testament, New York, Charles Scribner's Sons, 1965.

Hambly, W. Frank, Review of Calvin's Commentaries: The Gospel according to St. John and The First Epistle of John translated by T. H. L. Parker, 2 vol., 1959; Walter Lüthi: St. John's Gospel -- An Exposition translated by Kurt Schoenenberger, 1961; and The Structure of the Fourth Gospel by G. H. C. Macgregor and A. Q. Morton, 1961, in Australian Biblical Review 10:49-50, December, 1962.

Harrison, P. N., The Problem of the Pastoral Epistles, Oxford, Oxford University Press, 1921.

Herdan, Gustav, Language as Choice and Chance, Groningen, Holland, P. Noordhoff N. V., 1956.

Hinson, Glenn, Review of Morton and McLeman, Christianity in the Computer Age, in Review and Expositor 62:493-494, Fall, 1965.

Houlden, J. L., Review of The Structure of Luke and Acts by A. Q. Morton and G. H. C. Macgregor, in The Journal of Theological Studies, new series, 17:140-143, April, 1966.

IBM, Glossary for Information Processing, mim. (Available from IBM Corp., Armonk, N.Y.)

What is a Computer, Armonk, N.Y., IBM, 1965. (Available from IBM Corp., Armonk, N.Y.)

Jones, James L., "Computers and the Bible", in The Living Church, vol. 151, no. 15, pp. 14-15, October 10, 1965.

Le Corbeiller, Philippe, "What We Should Learn from Computers", in Proceedings of a Harvard Symposium on Digital Computers and Their Applications, pp. 1-6, Cambridge, Harvard University Press, 1962.

Leed, Jacob, editor, The Computer and Literary Style, Kent, Ohio, Kent State University Press, 1966.

Leicester, Ronald, Review of The Structure of Luke and Acts by A. Q. Morton and G. H. C. Macgregor, and Christianity and the Computer by A. Q. Morton and James McLeman, in The Churchman (London) 79:63-64, March, 1965.

B.6
5

Leicester, Ronald, Review of Paul, The Man and the Myth by A. Q. Morton and James McLeman, in The Churchman (London) 80:324-325, Winter, 1966.

Literary Data Processing Conference, Yorktown Heights, N.Y., 1964, Proceedings, edited by Jess B. Bessinger and others, New York, Materials Center, Modern Language Association, 1964.

Locke, William N., and Booth, A. Donald, editors, Machine Translation of Language, Cambridge, Technology Press, 1955.

McCasland, S. Vernon, Review of The Structure of the Fourth Gospel by G. H. C. Macgregor and A. Q. Morton, in The Journal of Biblical Literature 80:283, 1961.

Macgregor, G. H. C., and Morton, Andrew Q., The Structure of the Fourth Gospel, Edinburgh, Oliver and Boyd, 1961.

"Machines as Scholars' Helpers", in Educom 1:3-9, December, 1966.

Mitton, C. Leslie, The Epistle to the Ephesians, Oxford, Clarendon Press, 1951.

Review of The Structure of the Fourth Gospel by G. H. C. Macgregor and A. Q. Morton, in The Expository Times 72:365-366, September, 1961.

Morton, Andrew Q., and Macgregor, G. H. C., The Structure of Luke and Acts, London, Hodder and Stoughton, 1964.

Morton, Andrew Q., and McLeman, James, Christianity in the Computer Age, New York, Harper and Row, 1964. (Published in England as Christianity and the Computer.)

Paul, The Man and the Myth, New York, Harper and Row, 1966.

Mosteller, Frederick, and Wallace, David L., "Notes on an Authorship Problem", in Proceedings of a Harvard Symposium on Digital Computers and Their Applications, pp. 163-197, Cambridge, Harvard University Press, 1962.

Inference and Disputed Authorship: The Federalist, Reading, Mass., Addison-Wesley Publishing Co., 1964.

Nestle, Eberhard, editor, Novum Testamentum: Graece et Latine, 24n Auflage, novis curis elaboraverunt Erwin Nestle et Kurt Aland, Stuttgart, Württembergische Bibelanstalt, 1960.

"New Testament Studies: 7. Criticism by Computer", in The Hibbert Journal 63:43-44, Autumn, 1964.

"New Testament Studies: 9. How was Acts written?" in The Hibbert Journal 63:46-48, Autumn, 1964.

"New Testament Studies: 12. Paul and the Computer", in The Hibbert Journal 65:82, Winter, 1966-7.

Oettinger, Anthony G., Automatic Language Translation, Cambridge, Harvard University Press, 1960.

Parker-Rhodes, A. F., Review of The Structure of the Fourth Gospel by G. H. C. Macgregor and A. Q. Morton, in The Journal of Theological Studies, new series, 14:467-470, October, 1963.

Parrish, Stephen M., "Problems in the Making of Computer Concordances", Studies in Bibliography: Papers of the Bibliographic Society of the University of Virginia 15:1-14, 1962.

Rhys, J. Howard, Review of Christianity in the Computer Age by A. Q. Morton and J. McLeman, in Anglican Theological Review 48:118, January, 1966.

Robinson, Dow F., "Native Texts and Frequency Counts as Aids to the Translator", in The Bible Translator 14:63-71, April, 1963.

Satterthwait, Arnold C., "Programming Languages for Computational Linguistics", in Advances in Computers, vol. 7, pp. 209-238, edited by Franz L. Alt, New York, Academic Press, 1966.

Sebeok, Thomas A., "Linguistics Here and Now", in ACLS Newsletter 18:1-6, January-February, 1967.

Smalley, William A., "The Place of Linguistics in Bible Translation", in The Bible Translator 16:105-112, July, 1965.

Tasker, R. V. G., editor, The Greek New Testament: Being the Text Translated in The New English Bible, (1961), Oxford, Oxford University Press, 1964.

Tasman, Paul, "Literary Data Processing", in IBM Journal of Research and Development 1:249-256, July, 1957.

Vawter, Bruce, C.M., Review of The Structure of Luke and Acts by A. Q. Morton and G. H. C. Macgregor, in Interpretation 19:501, October, 1965.

Vincent, E. R., "Mechanical Aids for the Study of Language and Literary Style", in Literature and Science: Proceedings of the Sixth Triennial Congress, Oxford, 1954, pp. 56-60, International Federation for Modern Languages and Literatures, Oxford, Basil Blackwell, 1955.

Wake, William C., "The Authenticity of the Pauline Epistles", in The Hibbert Journal 47:50-55, October, 1948.

"Sentence-Length Distributions of Greek Authors", in Journal of the Royal Statistical Society, Series A, 120:331-346, 1957.

Williams, C. B., "A Note on the Statistical Analysis of Sentence-Length as a Criterion of Literary Style", in Biometrika 31:356-361, March, 1940.

"Studies in the History of Probability and Statistics IV. A Note on an Early Statistical Study of Literary Style", in Biometrika 43:248-256, December, 1956.

Winburne, John N., "Sentence Sequence in Discourse", in Proceedings of the Ninth International Congress of Linguists (Cambridge, Mass., 1962), pp. 1094-1099, The Hague, Mouton and Co., 1964.

"Toward a Structure of Discourse", typed. (1966)

Wisbey, Roy, "Concordance Making by Electronic Computer: Some Experiences with the 'Wiener Genesis'", in Modern Language Review 57:161-172, April, 1962.

Yule, G. Udney, "On Sentence-Length as a Statistical Characteristic of Style in Prose: With Application to Two Cases of Disputed Authorship", in Biometrika 30:363-390, January, 1939.

The Statistical Study of Literary Vocabulary, Cambridge, Cambridge University Press, 1944.